ACEEE

Proc. of Int. Conf. on Computational Intelligence and Information Technology

A Heuristic Approach for Network Data Clustering

Preeti Sharma[1], and Thaksen J. Parvat[2]
[1] Department of Computer Engineering, SKN SITS, Lonavala, India
Email: preeti5588@gmail.com
[2] Department of Computer Engineering, SIT, Lonavala, India
Email: pthaksen.sit@sinhgad.edu

***Abstract*— In this growing world of technology there are lots of security threats received by each and every area of computer networks. Most of the time the network security threats produce high false positive and negative ratios, this creates an obstacle for any security system to work improperly. The overwhelming threats make it challenging to understand and manage the network data.**

To address this problem we present a novel approach which eventually understand the network data by clustering them without background knowledge of any threats according to various parameters like source IP, Destination IP etc. And this approach saves administrator's time and energy in processing of large amount threats.

***Index Terms*— Threats, Alerts, Clusters, K- means, Email data, Enron, network security, email users, data mining.**

I. INTRODUCTION

Intrusions pose a serious security risk in a network environment. Network intrusion detection systems aim to identify attacks or malicious activity in a network with a high detection rate while maintaining a low false alarm rate. New emerging threats or attacks are the most difficult to detect. Signature based methods and misuse detection methods, which rely on labelled patterns, can detect previously known attacks with good accuracy but are unable to detect new types of attacks. In addition, maintaining the signature data base and labelling the patterns is time consuming and expensive.

Anomaly detection techniques can make use of un-supervised learning methods to identify new emerging threats with no need of labelled patterns, but, with a potential false alarm rate. We reviewed the different network intrusion detection methods and present here a comparative study with more emphasis on the unsupervised learning methods for anomaly detection.

The K-means algorithm was chosen to evaluate the performance of an unsupervised learning method for anomaly detection using the STES campus data set captured in the MS excel format at gateway. And also we used Enron Email data set to cluster emails. The results of the evaluation confirm that a high detection rate can be achieved while maintaining a low false alarm rate.

A network intrusion is any type of attack or malicious activity that can compromise the stability or security of a network environment. Intruders can be classified in two groups. External intruders don't have authorized access to the system or network they attack, while internal intruders have some authority access to the system. Network intrusions keep increasing over the years with new emerging and complex threats. This new emerging threats are the most difficult to identify.

A network intrusion detection system (NIDS) scans the network activities in a computer environment and attempt to detect the intrusions or attacks. Then, the sys-tem administrator may be alerted to the take the corrective actions.

DOI: 03.LSCS.2013.6.53

There are generally three types of approaches taken toward network intrusion detection. Signature-based,misuse detection and anomaly detection. The signature-based method is the oldest method in practice and depends on a signature database of previously known at-tacks. Misuse detection is a model-based supervised method which trains a classifier with labelled patterns to classify new unlabeled patterns. Anomaly detection approaches can make use of supervised or unsupervised methods to detect abnormal behaviours in patterns. A main objective of this study is to confirm the advantages of anomaly detection for intrusion detection using a simple clustering algorithm.

The rest of the paper is organized as follows: section 2 will introduce about related work done so far. Section 3 will give proposed work, Section 4 give result analysis process, section 5 will give future work and conclusion, Section 6 will give idea about implementation, and Section 7 References that we used in our proposed model.

II. RELATED WORK

TABLE I: LITERATURE SURVEY

Sl No.	IEEE Source/Year	Title	Purpose	Algorithm
1	ICET 2008 4th International Conference On Emerging Technology, 2008	An Intrusion Detection Mechanism Based on Feature Based Data Clustering	1.unsupervised anomaly detection has been used for Intrusion detection system (IDS) 2. identify attacks with a high detection rate and a low false alarm rate	Unsupervised k-means algorithm with labelling technique
2	4th IFIP International Conference on New Technologies, Mobility and Security (NTMS), 2011	Steps Towards Autonomous Network Security: Unsupervised Detection of Network Attacks	Detecting network attacks	temporal sliding-window algorithm
3	International Conference on Software Telecommunications and Computer Networks (SoftCOM), 2010	Anomaly Detection Using Baseline and K-Means Clustering	Anomaly Detection	K-means clustering and particle swarm optimization (PSO)
4	ICTAI 05, 17th IEEE International Conference On Tools With Artificial Intelligence, 2005	A Clustering Approach To Wireless Network Intrusion Detection	detecting intrusions or anomalous behaviour in WLANs	Online K-means
5	NSWCTC '09. International Conference on Networks Security, Wireless Communications and Trusted Computing, 2009	Study of Rough Set and Clustering Algorithm in Network Security Management	simplified network security assessment data set is established	Decision making rules, clustering by using rough set
6	IEEE 2nd International Conference on Software Engineering and Service Science (ICSESS), 2011	Analysis of Data Clustering Support for Service	routing anomaly detection in wireless sensor networks	Apriori and K-means algorithm
7	11th International Conference on Intelligent Systems Design and Application (ISDA), 2011	Efficient Data Clustering over Peer to Peer Network	clustering distributed databases	Distributed version of K-means algorithm
8	5th International Symposium on Telecommunications (IST), 2010	A New Hybrid Approach For Data Clustering	Proposed approach has suitable and acceptable efficacy in data clustering.	KAFSA(K-means and artificial fishes swarm algorithm)
9	IEEE-International Conference on Recent Trends in Information Technology, ICRTIT 2011	Data clustering with modified K-means algorithm	1. Decrease the complexity & the effort of numerical calculation 2.maintaining the easiness of implementing the k-mean algorithm. 3.assigns the data point to their appropriate class or cluster more effectively	modified K-means algorithm

Data clustering has been used in network area for different purposes like for intrusion detection, for analysis of support services etc which is depicted in table 1 [13].

Early work [2] utilized email traffic to infer social net-works for the purpose of discovering communities of shared interest. Email classification can be applied to several different applications, including filtering messages based on priority, assigning messages to user-created folders [5], or identifying SPAM.

One major consideration in the classification is that of how to represent the messages. Specifically, one must decide which features to use, and how to apply those features to the classification. [6] defined three types of features to consider in email: unstructured text, categorical text, and numeric data. Relationship data is another type of information that could be useful for Classification. Unstructured text in email consists of fields like the subject and body, which allow for natural language text of any kind. Generally, these fields have been used in classification using a bag-of-words approach, the same as with other kinds of text classification [3, 4 and 5].

A technique to classification the email is proposed by Martin, Sewani, Nelson, Chen, and Joseph [12]. The proposed classification is used for identifying spam messages. Categorical text includes fields such as "to" and "from". These differ from unstructured text fields in that the type of data which can be used in them very well defined. However, these fields have typically been treated the same as the unstructured text fields, with the components added to the bag of words [10, 11]. These fields have been found to be very useful in automatic email classification, although not as useful as the unstructured data [10, 12].

The characteristics of the links have been analyzed. Recently there are many analysis of Social Network [7]. In above-mentioned communication analysis, social network analysis techniques are often used. Social network analysis techniques are the methods to calculate the degree distribution of each node in networks, network density, and centrality.

Just as data mining can be loosely described as looking for patterns in data, text mining is about looking for patterns in text. However, the superficial similarity between the two conceals real differences. Data mining can be more fully characterized as the extraction of implicit, previously unknown, and potentially useful information from data [Witten and Frank, 2000].

The in-formation is implicit in the input data: it is hidden, un-known, and could hardly be extracted without recourse to automatic techniques of data mining. With text mining, however, the information to be extracted is clearly and explicitly stated in the text [4].

The Enron Email Dataset Database Schema and Brief Statistical Report [9] which shows how is the distribution of emails for every user and showing the network, how the employees are connected. Use of interpersonal communication for network inference has been of interest to researchers for several decades [1]. Early work [11] utilized email traffic to infer social networks for the purpose of discovering communities of shared interest. Email classification can be applied to several different applications, including filtering messages based on priority, assigning messages to user-created folders, or identifying SPAM.

III. PROPOSED METHOD

In this section, we describe our approach of clustering of network dataset and email dataset according to the steps shown in figure 1 and 3.

A. Network data clustering [13 and 14]

As shown in figure 1 there are 6 main steps in our approach of network data clustering that we describe the steps below

Step 1: In this step we have taken spread sheet of net-work data collected over the network of the STES campus by the main server at its end. This dataset is given as the input to our implemented system. Then our application with the help of jxl archive file it reads the spread sheet and it will take the input in the form of matrix.

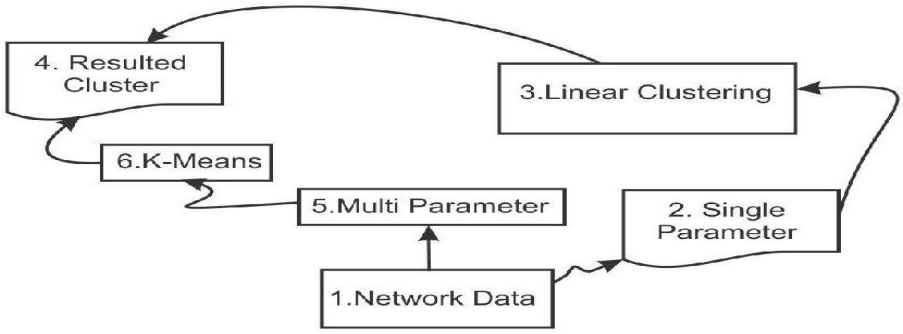

Fig 1: Network Data Clustering Approach

Step 2: In this steps user or administrator have to pro-vide a network parameter on the basis of which we need to cluster the network data. The parameters like Port number, Source IP, Destination IP etc.
Step 3: Here in this step the matrix of network data is been searched in its column for provided parameter name. If the parameter name is present in the column then for the column we search for the given parameter value. Then if we find the parameter value then we keep collecting the complete row value in a list. This operation keep iterating till the spread sheet completes. And that makes a cluster for the given value.
We propose the above Discussed Single parameter clustering using following Algorithm:
// input: Parameter name

1. **for** j=0 to columns
2. **if** (cell(i, j)=parameter name) **then**
3. **for** i=0 to rows
4. **if**(cell(i, j)=parameter value) **then**
5. Collect the row value into a vector to make a cluster

Repeat till the end of spread sheet to form cluster
Step 4: In this step collected cluster vector is written into a new spread sheet and saved in a given path.
Step 5: In this step the user or administrator has to pro-vide two network parameters on the basis of which we need to cluster the network data. The parameters like Port number and Source IP, time and destination IP etc.
Step 6: In this step the cluster is been formed using K-means algorithm for the multiple values.
K-means clustering is one of the unsupervised computational methods used to group similar objects in to smaller partitions called clusters so that similar objects are grouped together,. The algorithm aims to minimize the within cluster variance and maximize the intra cluster's variance. The technique involves determining the number of clusters at first and randomly assigning cluster centroids to each cluster from the whole datasets; this step is called initialization of cluster centroids [8].
The distance between each point in the whole dataset and every cluster centroid is then calculated using a distance metric. Then, for every data point, the mini-mum distance is determined and that point is assigned to the closest cluster. This step is called cluster assignment, and is repeated until all of the data points have been assigned to one of the clusters [8].
Finally, the mean for each cluster is calculated based on the accumulated values of points in each cluster and the number of points in that cluster. Those means are then assigned as new cluster centroids, and the process of finding distances between each point and the new centroids is repeated, where points are re-assigned to the new closest clusters. The process iterates for a fixed number of times, or until points in each cluster stop moving across to different clusters. This is called convergence. The steps of the K-means can be summarized in Figure 2 below.
The Algorithmic steps of the k-means can be summarized as below

- Initially, the number of clusters must be known, or chosen, to be K say.
- The initial step is to choose a set of K instances as centres of the clusters. Often chosen such that the points are mutually "farthest apart", in some way.
- Next, the algorithm considers each instance and assigns it to the cluster which is closest
- The cluster centroids are recalculated either after each instance assignment, or after the whole cycle of re-assignments.
- This process is iterated.

B. Email Data Clustering [13 and 14]

As shown in figure 3 there are 4 main steps in our approach. To cluster Email Data that we describe each and every step below
Step 1: In this step the collected Enron Email set folder is given as input to our system. Where Enron Email set is having folder with name of users and again within that it is having folder for inbox, outbox and sentbox emails.
Step 2: In this steps user or administrator have to pro-vide a parameter on the basis of which we need to cluster the Email data. The parameters are any word/s that we have to search within the emails. Again administrator have to provide searching folders like inbox, outbox and sentbox etc.
Step 3: Here in this step the given email folder item is scanned within each folder, the emails are read continuously and keep finding for the given string or word. If the given String is found then the user name is

collected in a vector. This operation keeps iterating till all users emails are scanned. And that makes a cluster for the given String or parameter.
We propose the above Discussed Email Clustering following Algorithm:
// input: Searching String
// input: Searching Email Folder like inbox, outbox, sentbox

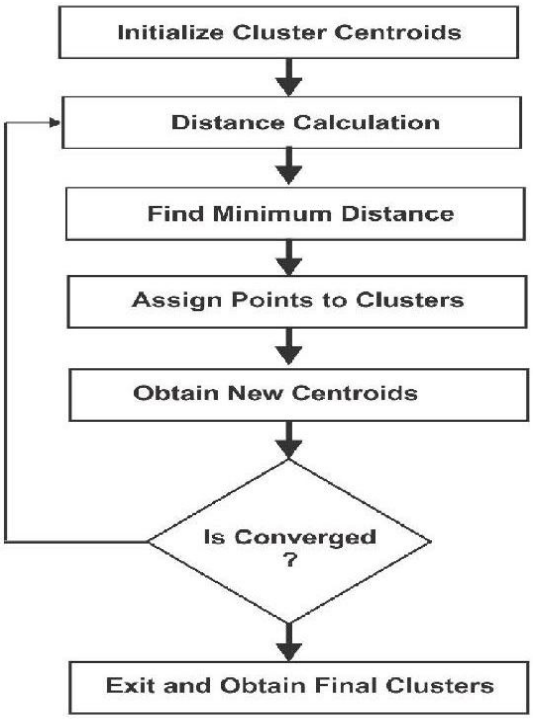

Fig 2: Computational steps of the k-means algorithm

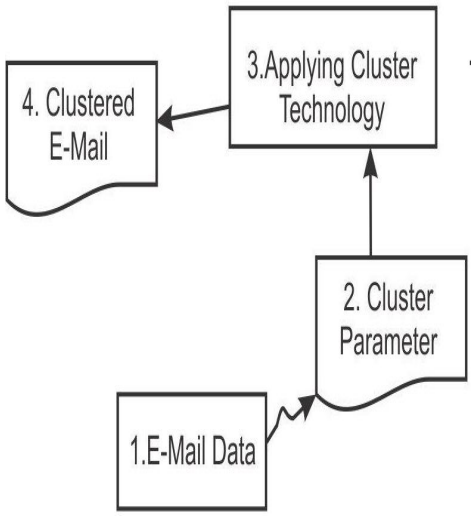

Fig 3: E-mail Data Clustering Approach

1. **for i** =0 to n //where n is number of User folders
2. **for j** =0 to m // where m is number of Email Folders
3. **if** (EmailFolder=Given Email Folder) **then**
4. **for** k=0 to P // where p is number of email files
5. Read complete Email and get into a String
6. **if** (Email string Contains searching word) **then**
7. get the User folder name into a vector
8. Repeat this for all User folders to get Cluster of users for specific word

IV. Results and Analysis

As we know Data clustering greatly boost the concept of grouping of similar objects. In our proposed and implemented system we apply this to cluster network data captured at server's end in the form of spread sheets, Where the spread sheet is having many columns like Source IP, Destination IP, Source Port etc,. The Result of these Clustering can shown and analyze below.

A. Result of Multiple Parameter Clustering

Here in the below table 2 we can see a resulted cluster ,which is for input of a multiple parameter for service and destination port Number value 27388 and 80 respectively, written in a spread sheet by our application in the same format as of given input spread sheet. This makes easy to analyze the data. The same kind of result we are getting in single parameter also, where we are giving only one parameter as input.

B. Performance Evaluation of Time For data clustering

Here in the below figure 5 we can see a performance of our system for formation of number clusters .If we ob-server the graph properly we can find that our system becomes lenient as the number of clusters increases. This means our system takes almost same time to form more number of clusters. This Indicates we will perform more accurately for huge amount data. This adds boosting point in our project.

C. K-Means Iteration Performance Evaluation

Here in the below figure 6 we can see that the number of iterations taken by the K-Means Algorithm to form clusters. If we observer Graph Properly we can see that K- means approximately takes almost same amount

TABLE II: OUTPUT OF MULTI-PARAMETER CLUSTERING

-------Cluster 0----------
For source and Destination Port [27388.0, 50983.0]
For source and Destination Port [27388.0, 41876.0]
For source and Destination Port [4771.0, 80.0]
-------Cluster 1----------
For source and Destination Port [27388.0, 443.0]
For source and Destination Port [56561.0, 8140.0]
For source and Destination Port [59201.0, 80.0]
For source and Destination Port [27388.0, 443.0]
For source and Destination Port [33885.0, 80.0]

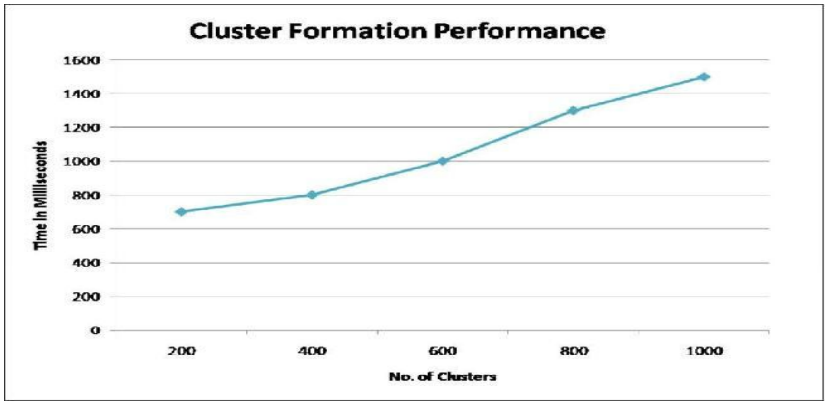

Figure 4: Cluster Formation Performance

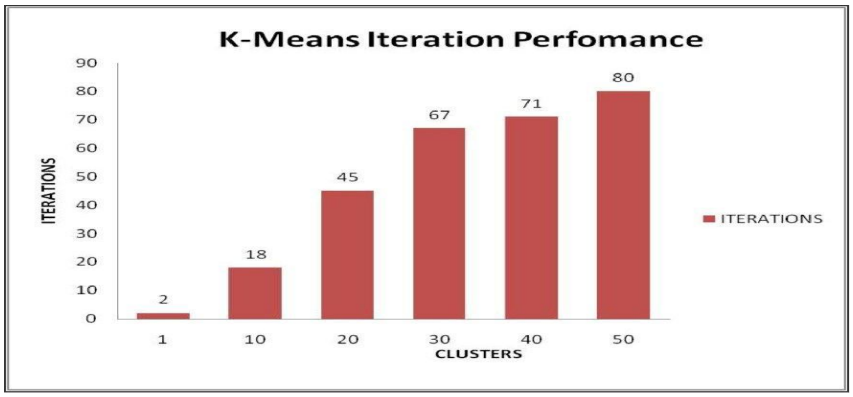

Figure 5: K-Means Iterations performance

of iteration to produce clusters. This shows that K-Means is having good performance ratio for greater amount of data to form cluster.

V. IMPLEMENTATION

The proposed algorithm is implemented using open source technologies and algorithm is applied over network data collected by the STES Lonavala campus server in the form of Microsoft excel sheets using Fortigate firewall and Enron email database respectively.

Java is selected as the programming languages and the other open source API's (Application Programming Interfaces) to Support the other functionalities. Net-Beans 6.9.1 is used as a development IDE (Integrated Development Environment) for Java and library of other technologies are added as external jar (Java Archives) in the Netbeans. NetBeans IDE is having feature of integrating the external libraries with it so that it can make easy to build any applications on its platform .jxl library available for Java that allows users to read and write the Microsoft spread sheets easily which eventually helps in the system to read data and to present clustered data.

CONCLUSION AND FUTURE WORK

In this paper network data clustering approach is pro-posed & implemented to show the clustered data which we can use for further studies of attacks & threats in a network. These clustered data can be used further to create an attack graph without the knowledge of previous attacks or threats. This eventually boosts the analysis of threats & alerts in network situation awareness. This kind of clustering systems can help the administrator to understand the network threats.

Here we also proposed & implemented a model for e-mail clustering to show text similarities. The proposed techniques shows the e-mail attributes & how the text similarities is use to cluster the users. The Enron dataset is used for the experiment in our project where the users are clustered based on many parameters in inbox, sentbox, drafts etc. folders of an email application.

The future scope of this work could be implementation of similar system for the e-mail attachments and real time e-mails of an organization. This kind of e-mail clustering helps in operations like summarization, automatic answering machines and conducting surveys.

REFERENCES

[1] Yingjie Zhou, Kenneth R. Fleischmann *"Automatic Text Analysis of Values in the Enron Email Dataset"* Proceedings of the 43rd Hawaii International Conference on System Sciences 2010

[2] Minoru Sasaki, Hiroyuki Shinnou" *Spam Detection Using Text Clustering"* Proceedings of the 2005 Inter-national Conference on Cyberworlds (CW'05) IEEE 2005

[3] Hui Huang, and Wei Jiang, Jason Cong, *"Pattern-Mining for Behavioral Synthesis"* IEEE, 2011

[4] Taiwo Ayodele, Shikun Zhou, Rinat Khusainov *"Evolving Email Clustering Method for Email Group-ing: A Machine Learning Approach"*, IEEE 2009

[5] Jayadev Gyani *"A Novel Approach for Clustering E-mail Users Using Pattern Matching"* Syeda Farha Shazmeen IEEE 2011

[6] Shady Shehata Fakhri Karray *"Enhancing Text Clustering using Concept-based Mining Model"* IEEE 2006

[7] Longbing Cao Huaifeng Zhang, Yanchang Zhao,Dan Luo, and Chengqi Zhang *"Combined Mining: Discovering Informative Knowledge in Complex Data"* IEEE 2011

[8] Data Clustering: A Review (1999) Jain/Murty/Flynn "http://citeseer.ist.psu.edu/jain99data.html"

[9] Brief Statistical Report, /"Technical Report, Information Sciences Institute, 2004. Available at: http://www.isi.edu/~adibi/Enron/Enron_Dataset_Report .pdf.

[10] W. W. Cohen: Learning Rules that classify E-mail. In Proc. of the 1996 AAAI Spring Symposium in In-formation Access, 1996.

[11] Y. Diao, H. Lu, and D. Wu "*A comparative study of classification based personal email Filtering*" In Proc. 4th Pacific-Asia Conf. Knowledge Discovery and Data Mining (PAKDD'00), pages 408-419, Kyoto, JP, 2000.

[12] Steve Martin, Anil Sewani, Blaine Nelson, Karl Chen, Anthony D. Joseph, "*Analyzing Behaviorial Features for Email Classification",* Second Conference on Email and Anti-Spam CEAS 2005 , The International Association for Cryptologic Research and The IEEE Technical Committee on Security and Privacy 2005.

[13] Sharma P., Parvat T. J. *"Network log Clustering Using K-means Algorithm"* pp 97-105, ITC 2012, Springer-Verlag Berlin Heidelberg 2012 Jiawei Han, Micheline kamber.*: Data Mining Concepts and techniques*.: Second Edition, Morgan Kaufmann Publishers (2006)